α-Tweet: Predicting retweets

Souradeep Sinha
Department of Computer Science
Syracuse University
ssinha04@syr.edu

ABSTRACT

The act of retweeting has always been an important measure of knowledge propagation ever since microblogging was popularized by Twitter in 2006. It has always been a hard task to measure the chances of a particular tweet getting retweeted, before actually tweeting. This paper uses Naïve Bayesian Classification to predict the possible number of times the tweet in question would get retweeted, by tracking the retweets by the followers of any given user.

CCS Concepts

• **Information systems → Database management:** Data Mining

Keywords

Tweet popularity prediction, Naïve Bayesian Classification, Social Media

1. INTRODUCTION

Twitter revolutionized the manner in which human beings participated in online social platforms. Not only did the 140 character limit on content made it easy to read and absorb shared information, it also gave rise to an art of trying to make the most sense out of limited written text. With art, came appreciation – as views, ideologies, opinions, motives and of course raw facts were either accepted wholeheartedly or debunked drastically with equal passion. This universal appeal of Twitter attracted stalwarts from Planet Earth to use it as a channel in order to get their message across to the online population. And suddenly, it became important to estimate and make sure that their quota of 140-charactered tweet to the world penetrated wide, and deep. Enter retweets, a Twitter owned concept of chain propagation of messages by a user's followers for the "next hop" followers to view, and recursively retweet even further.

There is a plethora of commercial and Twitter owned analytical tools for Twitter data to arm a user with variety of reports and visualizations to help with content management and tracking trends among followers. However, predicting the "retweetability" has been so far mostly an unsolvable quest, and has been tried and tested with different methodologies. Because of the uncertain nature of a human being's behavior on social media, it works as a convenience if it was indeed possible to know how one's thoughts will fare among the "Tweeple", and more importantly, how far the outreach is going to be. This is especially true for firms who are responsible for managing the publicity and online presence of distinguished personalities.

This project aims at trying to tackle the challenging problem, and testing whether the approach provides a solution that could be advanced with further research. Even though retweet predictions are difficult by itself, because of the involvement of societal sentiments as a whole, it is further worsened by the fact that currency of issues play a major role in understanding popularity. A meticulous survey and case study was undertaken by (Boyd, et al., 2010), who documented the act of retweeting as a participant to participant conversation. (Zaman, et al., 2010) train a probabilistic collective filtering model on datasets of username and corresponding retweets to predict future retweets. they also argue that the identity of the original tweeter assumed a crux role in garnering retweets. (Suh, et al., 2010) identified that URLs and hashtags, as well as age of user and number of followers account for a strong relationship with retweetability, while the number of past tweets do not have any effect on the prediction of retweets while using a predictive retweet model. (Nagarajan, et al., 2010) use link-based diffusion models to make observations on event driven tweets and find a distinct relationship between sparse and dense retweet patterns with the content and tweet type. The closest attempt to predicting retweets comes from (Petrovic, et al., 2011), who use a passive-aggressive machine learning algorithm, and found that social features dominated the learning.

The main motivation of this paper came from actual conversations with Twitter users, most of whom had the view that the chances of a given tweet getting retweeted depends heavily on the content that the followers have been retweeting mostly about in the recent past. Apart from this highly preferred answer, another general notion is also how the individual's social standing to her/his followers are, and with what weight that the causes the individual stands for that gets retweeted further in the current past also have a significant effect.

Since currency of events hold a significance according to our theories, we make sure that we only collect data for the last four weeks for both the users and followers. Our model takes an input of the username and the text that the user wants to tweet. Our algorithm then gives a possible integer estimate of how many possible retweets that exact text is going to garner. We track the "hotness" of topics present in the subject tweet among the followers and use an accumulating product with the retweet popularity of those topics within the user's current tweets. The general approach is to compare this accumulated product with the accumulated product of not getting a retweet per follower, but because the number of retweets is generally much lesser compared to the number of followers, we choose a small fraction arbitrarily as the threshold. Once the fraction crosses that threshold for the particular follower and count it as a likely retweeter. In the end, we return the total number of such possible retweeters.

2. FURTHER RELATED WORK

The notion of retweeting can be extended to virality in other forms of blogging, and pioneering work was done by (Hoang, et al., 2012) and use the concepts of user susceptibility, topic virality and user virality to model a set of observed retweet data. However, their proposed V2S model is limited to the time range of a certain event, in this case the elections of Singapore.

Apart from the works of researchers already discussed above, (Naveed, et al., 2011) use content-based features to train a model that contributed to finding the likelihood of a retweet. Using content analysis such as language used in tweet, user's social network presence and history, (Artzi, et al., 2012) train models to with real world data to predict if a tweet will receive a response in the form of a retweet or reply.

3. QUESTION: RETWEET OR NOT?

Retweeting, can be differentiated from replying because it subverts the realm of one to one communication, and it not just re-posts a tweet, but for a follower as an independent user of Twitter, exhibits their alignment towards a given topic. For the *retweetee*, it signifies the acceptance of his/her views by the people who "follow" him/her (Fouts, 2011). Given such importance being associated to retweeting, it becomes convenient for a user to know in advance how his/her tweet is going to fare in terms of number of retweets. We formally define this problem of predicting number of retweets as follows:

*Given a test tweet **t** from a user **U** with his collection of tweets over a time period of 4 weeks **T**, with set of followers **F**, and set of followers' retweets, $\boldsymbol{R_f}$ where $f \in F$, this paper aims to find the number of retweets **n** for the tweet t.*

4. α-TWEET: RETWEET PREDICTOR

Retweeting, can be differentiated from replying because it subverts the realm of one to one communication, and it not just re-posts a tweet, but for a follower as an independent user of Twitter, exhibits his/her quality of affiliation. Hence, in a manner of saying, the number of retweets is a good way of gauging the popularity on the user's part.

A few considerations that were made while designing α-TWEET are as follows:

4.1 Currency of information

As observed by (Hoang, et al., 2012), this project tries to preserve the time associated virality property of the topics by making sure that the tweets of the user and followers are within the last 28 days.

4.2 Identity related topical favorability

When the user tweets, it also shows a lot about the users affiliation to certain topics, and how likely he is to get a retweet based on the result of this affiliation. We determine that by taking a fraction of the sum of the number of times tweets by the user containing a specific topic have been retweeted, over the number of times the user has been retweeted in total.

4.3 Topic related follower favorability

To capture the essence of "fan following" that the user has, we get a fraction of all the times the followers have retweeted a certain topic over the number of times they have retweeted in total over the last 28 days.

4.4 User's real life popularity

In accordance to findings of (Suh, et al., 2010), who believe that the identity, age and number of followers have a correlation to the number of times they get retweeted. For this project four very successful and well known comedians based out of USA, and their tweets and followers are handpicked to be used for training the models.

5. PROCEDURE

In this section, we systematically talk about the data, the procedure used to design, and performance evaluation.

5.1 Data Collection

No preexisting datasets were used for this project. In fact, for the chosen user, a scraper was written which would collect all the tweets for the input user, and the retweets of all his/her followers and save them in csv format in a separate folder. All data collected were over a period of 28 days. The data was then divided into two sets. The test set consisted of user's tweet for the last seven days, while the training set consisted of user's as well as followers' tweets and retweets respectively over the next 21 days. We reject the retweet data of the followers for the last seven days as they do not play a purpose in testing the model.

The data collected was also strictly in English, and ones whose <language> tag was not "en" were rejected. We also reject followers who do not have any retweets over the given frame of time because of their inactivity.

An alternative that could have been used for this project would be to use Twitter's REST APIs. However, because of its limitation of use as to the maximum number of API calls over 15 minute time slots, we were forced to use the scraper. Though the scraper is unbounded, it uses Selenium driver[1] to open up browsing windows to sift through multiple screens and copy source code. A downside to this technique entailed that collection of data was slow, and given our choice of celebrities with huge number of followers, would take a lot of time. To overcome this problem, we precollected data for 4 well known comedians on December 12, 2015, and included all information up to 28 days prior.

[1] http://selenium-python.readthedocs.org/

Table 1. Data used for building model

U	$\|\|F\|\|$	$\|\|T\|\|$	$\sum\|\|R_f\|\|$
@mulaney	2906	17	70188
@azizansari	515	10	7262
@amyschumer	895	65	14224
@therealrusselp	596	23	7857

5.2 Data Format

All of the related data is stored in a folder titled as /<username>. The user's tweet is stored as a CSV file as u_<username>.csv and his/her follower's retweet history is stored as f_<follower>.csv for all his followers.

Data was stored in the following format in each row, regardless of user or follower, tweet or retweet:

<timestamp>, <text of tweet>, <number of RT>, <number of favorites>

Or

```
<timestamp>, <text>, <num_RT>, <num_fav>
```

For the user, all of his/her retweets were ignored, while for the follower, all of their tweets were ignored and retweets were considered.

5.3 Data Cleaning and Preprocessing

Apart from filtering the aforementioned data elements, the tweet text was cleaned as well. All special characters, except @ and # were removed. This was to make sure that we preserved the hashtags and mentions, which are two very important aspects of Twitter. The NLTK stop word corpus for English (NLTK Project, 2015) were used to remove all words that serve as grammatical fillers and would have no information content. Removing the dots (.) from the text meant that all hyperlinks were destroyed, however, the name of the parent website was conserved, and if the same URL was being used in multiple tweets, the broken hyperlink parts still remained intact in all cases.

5.4 Algorithm

For every word of the input tweet from a given user, we aim to find the probability of that words contribution to a retweet. We hope to achieve this by taking a product of the word's identity related topical favorability (IRTF) discussed in section 4.2 with the word's topic related follower favorability (TRFF) discussed in section 4.3.

For a word, we calculate IRTF as the fraction of retweets that the tweets containing that word has received over all retweets that the user has received.

Calculation of TRFF is almost similar. For each follower, we find the ratio of how many times that the follower has retweeted a tweet containing the word over the total number of times the follower has retweeted.

Once we have the $P(t) = \sum IRTF(w_i) * TRFF(w_i)$ over all w_i in input tweet, t, we check that for all of the followers in F. The predicted number of retweets is the number of followers whose P(t) values exceed 0.001.

We take the arbitrarily small value of 0.001, as a threshold because the number of followers of distinguished personalities is generally much higher than the number of retweets he/she gets per tweet.

We define two functions classify() and Driver() as follows[2]:

```
function classify(U, t, T, $R_f$):
  total_rt ← 0
  for tweet ut in T:
    total_rt += ut.<num_RT>
  end for
  P_ $w_i$ ← {}
  for word $w_i$ in t:
    $w_i$_rt_count ← 0
    for tweet ut in T:
      if $w_i$ in ut:
        $w_i$_rt_count += ut.<num_RT>
      end if
    end for
    P_$w_i$[$w_i$] ← $w_i$_rt_count / total_rt
  end for
  P_$w_i$ ← vectorize P_$w_i$
  follower_vectors ← []
  for f in F:
    f_P_ $w_i$ ← {}
    for word $w_i$ in t:
      $w_i$_ft_count ← 0
      for retweet ft in $R_f$[f]:
        if $w_i$ in ft:
          $w_i$_ft_count ++
        end if
      end for
      f_P_$w_i$[$w_i$] ← $w_i$_rt_count / len($R_f$[f])
    end for
    follower_vectors.append(vectorize f_P_$w_i$)
  end for
  count ← 0
```

[2] This pseudocode does not cover the data cleaning and preprocessing portion.

```
  for v in follower_vectors:
    bayes_prob ← 0
    for i in v:
      bayes_prob += v[i] * P_$w_i$[i]
    end for
    if bayes_prob > 0.001:
      count++
    end if
  end for
  return count
end function

function Driver():
  U ← Username as input
  T_full ← Collect the user’s tweets
  T_test ← Tweets in last 7 days
  T_train ← Tweets in (7,28] days
  F ← Get all his followers
  $R_f$ ← {}
  For all f in F:
    $R_f$[f] ← f’s retweets in (7, 28] days
  end for
  t ← Tweet as input
  Clean t and remove stopwords
  predicted_rt = classify(U, t, T_train, $R_f$)
  output → predicted_rt
  test_predicted_RTs ← []
  test_actual_RTs ← []
  for test_t in T_test:
    test_actual_RTs.append( test_t.num_RT)
    test_predicted_RTs.append( classify(U, test_t, T_train, $R_f$)
  end for
  output ← Pearson_Corr( test_actual_RTs, test_predicted_RTs)
```

```
end function
```

6. RESULTS AND DISCUSSIONS

To check our classification task, we find out 4 tweets, 1 per each user that is not present in the training set or testing set. These are t1, t2, t3 and t4 and tweeted by @mulaney, @azizansari, @amyschumer and @therealrussellp respectively. t2 was taken from Oct 17 because the user did not actually tweet after the last tweet was captured by the training set. Care was taken that the tweet was not influenced by too many event driven words to avoid the effect of recency of tweet to a particular time frame.

t1: Alda News That's Fit to Print. The New York Times' Ben Brantley on a tuna based play written by elderly lunatics. http://nyti.ms/1jRWgyc

t2: Oddball fest is done! Thanks everyone who came to the shows.

t3: Nashville is the shit with @marknorm and @Steinbassclar and #locksmithisadore

t4: So this happened last night!!! Went to the Premier of the newest StarWars... What a great after… https://www.instagram.com/p/_U4_f5oNJV/

We fed these tweets to our classifier, and the noted the Pearson's correlation coefficient to determine the performance of the classifier. The following were the results in Table 2 and Table 3.

Table 2. Predicted vs Actual number of RTs

User	Input tweet	Predict RTs	Actual RTs
@mulaney	t1	13	36
@azizansari	t2	43	512
@amyschumer	t3	0	70
@therealrussellp	t4	3	188

Table 3. Pearson's Correlation Cefficient

User	Input tweet	Corr	p-value
@mulaney	t1	-0.914	0.08
@azizansari	t2	-1.0	0
@amyschumer	t3	-0.179	0.506
@therealrussellp	t4	-0.546	0.34

Given the negative correlation values, the rusts are surprising because it implies that the predicted values are, for some reason being oppositely affected by the words present in the input tweet when tallied against the TRFF and IRTF measures described above. This behavior is also unintuitive because we cannot think of possible causes that could have affected this opposite behavior.

There is a vast scope for further work on this project as majority of the concepts were developed with the help of extensive related research in the field already been done previously by known researchers. It may be advisable to probably test out the impact of sentiment on the classifier by using a sentiment measurement tool. Another possible feature may be to find out how the different moods surrounding the input tweet may be interpreted by the followers and that could reversely affect the way the number of retweets are predicted.

7. REFERENCES

“Tweet, Tweet, Retweet: **[Conference] / auth. Boyd Danah, Golder Scott and Lotan Gilad // IEEE. - Kauai, HI : [s.n.], 2010.**

A Qualitative Examination of Topical Tweet and Retweet Practices [Conference] / auth. Nagarajan Meenakshi, Purohit Hemant and Sheth Amit // International Conference on Weblogs and Social Media (ICWSM). - Washington, DC : Kho.e.sis Publications, 2010.

Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter [Conference] / auth. Naveed Nasir [et al.] // WebSci: Proceedings of the ACM. - Koblenz : [s.n.], 2011.

Natural Language Toolkit [Online] / auth. NLTK Project // NLTK Corpora. - 2015. - http://www.nltk.org/nltk_data/.

Predicting Information Spreading in Twitter [Conference] / auth. Zaman Tauhid R [et al.] // Neural Information Processing Systems. - Vancouver : [s.n.], 2010.

Predicting Responses to Microblog Posts [Conference] / auth. Artzi Yoav, Pantel Patrick and Gamon Michael // Association for Computational Linguistics: Human Language Technologies. - Montreal : [s.n.], 2012.

RT to Win! Predicting Message Propagation in Twitter [Conference] / auth. Petrovic Sasa, Osborne Miles and Lavrenko Victor // ICWSM. - Barcelona : [s.n.], 2011.

The Social Media Coach [Online] / auth. Fouts Janet // The Art of Retweet. - 2011. - http://janetfouts.com/the-art-of-the-retweet/.

Virality and Susceptibility in Information Diffusions [Conference] / auth. Hoang Tuann-Anh and Lim Ee-Peng // ICWSM. - London : [s.n.], 2012.

Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network [Conference] / auth. Suh Bongwon [et al.] // IEEE Second International Conference on Social Computing (SocialCom). - Minneapolis, MN : IEEE Computer Society, 2010.